HW 3

Yasaman, Firooz, and Ozan.

$\boxed{\text{Notation: } \langle x, y \rangle = x^T y}$ ⟵ Inner product.

① $\min_{w} f(w)$ s.t $Aw - b = 0$

$$g(\lambda) = \min_{w} f(w) + \langle \lambda, Aw - b \rangle$$

$$g(\bar\lambda) = \min_{z} f(z) + \langle \bar\lambda, Az - b \rangle$$

$$\leq f(w) + \langle \bar\lambda, Aw - b \rangle$$

$$= f(w) + \langle \bar\lambda - \lambda, Aw - b \rangle + \langle \lambda, Aw - b \rangle$$

$$= g(\lambda) + \langle \bar\lambda - \lambda, Aw - b \rangle$$

⟶ for this equality we need $w \in W(\lambda)$ where

$$W(\lambda) = \arg\min_{w} f(w) + \langle \lambda, Aw - b \rangle$$

↓ set of solutions

sub-gradient equation for convex function f: $f(y) \geq f(x) + \langle \partial f(x), y - x \rangle$

we have $g(\bar\lambda) \leq g(\lambda) + \langle \bar\lambda - \lambda, Aw - b \rangle$

$$\longrightarrow \quad Aw - b \in \partial g(y)$$

2 $\max_{\lambda} \min_{\omega} \underbrace{f(\omega) + \langle \lambda, A\omega - b \rangle}_{L(\omega,\lambda)}$

Dual ascent $\longrightarrow$ do Alternating Minimization.

$$\begin{cases} \omega_{k+1} \in \arg\min_{\omega} L(\omega, \lambda_k) \\ \lambda_{k+1} = \lambda_k + \alpha_k \left(A\omega_{k+1} - b \right) \end{cases}$$

f is L-smooth and μ strongly-cvx

$\Longrightarrow$ $L(\omega,\lambda)$ is also L-smooth and μ-strongly convex since we are adding only a Linear term.

$\Longrightarrow$ The problem has a unique solution.

Regarding rate of convergence for primal variable update if you use GD with fixed step-size:

It converges at Linear rat $O\left(\log(1/\varepsilon)\right)$

ε-optimal solution.

Regarding rate of convergence for the dual varible:

Consider that we are using Gradient Ascent, and similarly we will have $O(\log(1/\varepsilon))$ rate of convergence, if we show the dual function is both smooth and strongly convex:

Dual function $\longrightarrow$ $\underbrace{f^*(\cdot)}$ + $\underbrace{\lambda^T b}_{\text{Linear}}$ $\longrightarrow$ does not change strong convexity and smoothness parameters.

We know from properties of conjugate function:

if $f(\cdot)$ is L-smooth and μ strongly convex $\longrightarrow$ f^* is $1/\mu$ strongly convex and. $1/L$ smooth.

So, the total rate of convergence is $O(\log(1/\varepsilon))$.

checking primal solution is feasible: Fixed point Analysis: $\longrightarrow$ $\omega_{k+1} = \omega_k = \omega$

$\lambda_{k+1} = \lambda_k = \lambda$

fixed-point iteration

$$\omega = \underset{x}{\arg\min}\ L(x, \lambda) = \underset{x}{\arg\min}\ f(x) + \langle \lambda, Ax - b \rangle$$

$$\lambda = \lambda + \alpha_k (A\omega - b) = 0 \implies A\omega = b$$

so if the algorithm converge, it will converge to primal feasible.

Second interpretation:

The problem is convex with equality constraints. $\longrightarrow$ slater's condition holds

$\longrightarrow$ The solution is primal feasible and optimal.

③

$$\min \; \underbrace{\frac{1}{N}\sum_{i=1}^{N} f_i(\omega_i)}_{F(W)}$$

s.t $\boxed{\omega_i = w_j \quad \forall j \in N_i}$ is equivalent to $W\sqrt{R} = 0$ where

$W = [w_1 \; w_2 \dots w_N] \in \mathbb{R}^{d \times N}$

R: gossip matrix $\in \mathbb{R}^{N \times N}$

same sparsity pattern as graph Laplacian.

R is positive semi-difinite

$\longrightarrow$ $\sqrt{R}$ exists, and if $R = V^T \Sigma V$, then:

$$\sqrt{R} = V^T \Sigma^{1/2} V$$

$R = D - A$

D: degree matrix.

A: adjecancy matrix.

By defining $F(w) \triangleq \frac{1}{N}\sum_{i=1}^{N} f_i(\omega_i)$, problem is equivalent to:

$$\min_{w \in \mathbb{R}^{d \times N}} F(W) \quad \text{s.t } W\sqrt{R} = 0 \quad \xrightarrow[\text{problem}]{\text{Dual}} \quad \max_{\lambda \in \mathbb{R}^{d \times N}} -F^*(\lambda \sqrt{W})$$

F^* is conjugate of F

To solve this maximization $\longrightarrow$ $\lambda_{t+1} = \lambda_t - \eta \nabla F^*(\lambda_t \sqrt{W}) \sqrt{W}$

change of variable $y_t = \lambda_t \sqrt{W}$

$$\longrightarrow \quad y_{t+1} = \underbrace{y_t - \eta \nabla F^*(y_t) W}_{\text{similar to gossiping } \nabla f_i^*(y_{i,t})}$$

So the final algorithm works as

1. Initialization
2. For iterations t:
3. each node computes its $\nabla f_i^*(y_{i,t})$ $\longrightarrow$ each node does its local computations.
4. sends ∇f_i^* to its out neighbors.
5. update $y_{i,t+1} = y_{i,t} - \eta \sum_{j \in N_i} r_{ij} \nabla f_j^*(y_{j,t})$ $\longrightarrow$ update dual variable.
6. end for.

* primal distributed method $\omega_{i,t+1} = \sum_{j \in N_i} r_{ij}\, \omega_{j,t} - \mu \nabla f_i(\omega_{i,t})$ does not converg to the exact

solution of problem $\min_{s.t\ w_i = w_j} \frac{1}{N} \sum_{i=1}^{N} f_i(w_i)$, and there exists a distributed variance.

to make it clear check matrix Notation: $W = RW - \mu \nabla F$

fixed point if reaches consensus $\nabla F = \begin{bmatrix} \nabla f_1 \\ \nabla f_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \end{bmatrix} \Rightarrow \nabla f_i = 0$, instead optimal solution $\sum \nabla f_i = 0$

$\Longrightarrow$ This algorithm does not converge to optimal solution

$\longrightarrow$ But the dual method will converge.

$\longrightarrow$ Regarding Communication in each iteration,

Both methods Communicate the same number of bits.